JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 3, No. 3 June 2021 DOI: https://doi.org/10.34288/jri.v3i3.68

SENTIMENT ANALYSIS OF TWITTER DATA ON DISTANCE LEARNING USING NAÏVE BAYES ALGORITHM

Putri Rana Khairina [1*], Desti Fitriati[2]

Informatics Engineering Study Program[1,2]
Pancasila University[1,2]
putrirnkh@gmail.com[1], desti.fitriati@univpancasila.ac.id[2]
(*) Corresponding Auhtor

Abstrak

Covid-19 menyebar secara luas hingga mengakibatkan pandemi global. Sistem Pembelajaran Jarak Jauh (PJJ) dianggap sebagai solusi tetapi, kenyataan saat pelaksanaan PJJ ini belum sesuai dengan harapan dari masyarakat. Pengguna twitter banyak menuliskan pendapatnya terhadap PJJ. Kecenderungan sentimen masyarakat dapat digunakan sebagai salah satu cara membenahi sistem pendidikan yang ada di Indonesia dan dapat menjadi masukan bagi pemerintah dalam menyempurnakan metode PJJ yang sedang dilaksanakan. Maka, penelitian ini menghasilkan sebuah sistem yang dapat menganalisis sentimen *tweet* terhadap PJJ. *Tweet* tersebut didapat menggunakan Twitter API. Metode yang digunakan adalah *Naïve Bayes* untuk proses klasifikasi sentimen positif, negatif dan netral dengan menggunakan 600 data. Kemudian, dilakukan pembagian data 80% data *training* dan 20% data *testing* yang akan di *text preprocessing* terlebih dahulu. Tingkat akurasi analisis sentimen terhadap PJJ dengan metode *Naïve Bayes* menggunakan 3-*fold Cross Validation* menghasilkan rata – rata sebesar 93%.

Kata kunci: *Covid-19, Pembelajaran Jarak Jauh (PJJ), Naïve Bayes, Analisis Sentimen, k-Fold Cross Validation.*

Abstract

Covid-19 is widespread, resulting in a global pandemic. Distance Learning System (DLS) is considered as a solution but, the reality of the implementation of DLS is not in accordance with the expectations of the community. Many twitter users wrote their opinions on DLS. The tendency of public sentiment can be used as a way to improve the existing education system in Indonesia and can be an input for the government in improving the DLS method that is being implemented. Thus, this study produced a system that can analyze tweet sentiment towards DLS. The tweet was obtained using the Twitter API. The method used is Naïve Bayes for the process of classification of positive, negative and neutral sentiments using 600 data. Then, data sharing is done 80% data training and 20% data testing that will be in the text preprocessing first. The accuracy of sentiment analysis of DLS using Naïve Bayes method using 3-fold Cross Validation produces an average of 93%.

Keywords: Covid-19, Distance Learning System (DLS), Naïve Bayes, Sentiment Analysis, k-Fold Cross Validation

INTRODUCTION

Covid-19 is a new type of corona virus found in Wuhan, Hubei, China in 2019, named Coronavirus disease-2019 which is shortened to Covid-19. Covid-19 has since been found to have spread widely, resulting in a global pandemic. Corona Virus is a type of virus that can cause mild symptomatic illness to severe symptomatic illness and the worst impact is death(Siagian, 2020).

This virus is very quickly transmitted to almost all countries, including Indonesia. The effect created by the Covid-19 pandemic has assaulted the wellbeing area, yet additionally assaulted different areas, for example, the travel industry, account, business, to transportation (Salam, 2020). At this time all actvities are carried out at home, with the aim of tackling the increasing number of Covid-19 sufferers in Indonesia. At present, the spread of Covid-19 can occur through air, surfaces contaminated by viruses, droplets and through human waste (CAKTI INDRA GUNAWAN & YULITA, 2020). One form of anticipation for Covid-19 countermeasures, the Government issued apolicy Social Distancing and Physical Distancing. One form of government countermeasures, namely Large Scale Social Restrictions (LSSR) for all activities that meet directly with other people, one of which is learning activities. Therefore, the education sector implements Distance Learning System (DLS) or online learning which is carried out at home.

The Ministry of Education and Culture as the provider of national education issued a circular from the Minister of Education and Culture Number 3 of 2020 which contains the Prevention of Covid-19 in education, and circular letter Number 4 of 2020 which contains the implementation of Education Policy in an Emergency for the Spread of Covid-19(Ahmad, 2020). This step is taken to reduce or minimize the number of patients exposed to the virus.

The form of equal distribution of education, such as during the Covid-19 pandemic, is to prepare technology to help teachers provide learning materials such as Application Zoom, Google Meet, and so on. All education sectors, such as; Early childhood, kindergarten, elementary, junior high school, up to universities that previously did teaching methods face to face when learning took place, now all converted done online.

This Distance Learning System (DLS) is considered a solution that can be used as a means of equitable education without having to face to face to avoid the spread of the virus. However, the reality at the time of implementing this DLS was not in accordance with the expectations of the community. This is the public sentiment towards distance learning methods on social media twitter (Sulaiman, 2020).

Twitter is one of the social media that is widely used by the community, and allows its users to send and read text-based messages. Twitter users write opinions or opinions on various topics. Users can express about a variety of topics usually all issues that are being discussed from various categories of life, such as; Entertainment, politics, social, and government are busy being discussed here (Rosyad, 2019). At this time, Twitter was busy talking about Distance Learning System (DLS). Twitter users write a lot of their opinions about DLS, which can be in the form of support, input, complaints, criticisms that can be found in the form of writing. These sentiments represent the feelings and emotions that people have towards DLS. The tendency of public sentiment can be used as a way to fix the education system in Indonesia. Twitter data can be used as a source of research data, because Twitter has short messages called tweets that contain various opinions or opinions of the public, and fast news distribution. In addition, these tweets will also continue to add up fast in realtime. So, a sentiment analysis system is needed to analyze the opinions of the community that are constantly changing rapidly and can be used as a benchmark for the government to procure the DLS method.

Sentiment analysis or opinion mining will classify a person's opinions, sentiments, evaluations, attitudes, and emotions into written language to find out opinions in the form of positive, negative or neutral sentences (Risnantoyo et al., 2020). In classifying a sentence or tweet, many methods can be used, one of which is themethod Naïve Bayes. Naïve Bayes is used to find the highest probability value for classifying twitter data. The advantage of Naïve Bayes only requires a small amount of training data to find the parameters of the required variables and is independent of each other (Devita et al., 2018).

According to research by Muhammad Husni Rifqo and Ardi Wijaya (Rifqo & Wijaya, 2017), this researcher conducted research on "Implementation of theAlgorithm Naïve Bayes in Determining Credit Lending". The data used is the ACC company credit Agiing Data Set from 2010 to 2011 and uses the existing data set in the UCI. This study produces good accuracy because the number of data sets affects the level of accuracy.

According to the research of Tutus Praningki and Indra Budi (Praningki & Budi, 2018), this researcher conducted the "Cervical Cancer Prediction System using CART, Naïve Bayes, and k-NN". The data used is RSUD Kediri with a total of 702 records. The results of this study show the classification accuracy with the CART,algorithms Naïve Bayes, and k-NN. Thealgorithm Naïve Bayes performs classification with the highest accuracy rate of 94.44%. When compared with the CART Algorithm, the accuracy is 88.89%, and the k-NN Algorithm produces an accuracy of 85.04%.

Based on previous research using themethod Naïve Bayes has a high degree of accuracy so that it can make the resulting data more valid. So, in this study, the aim of this research is to create a system of public sentiment analysis, especially Twitter users, regarding DLS during the Covid-19 pandemic. This system is designed by classifying into 3 forms of data, namely; Positive, Negative, and Neutral. Where in it contains data conclusions from public sentiment towards DLS in the form of tweets. This sentiment analysis system can provide input for the government in perfecting the DLS method that is being implemented and can fix the existing education system in Indonesia.

RESEARCH METHODS

Needs Analysis

a. Functional Analysis

The functions of this system are as follows:

- Enter data on the data training menu.
- Displaying data training and data testing.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v3i3.68

- Displaying results preprocessing on data training and data testing.
- Displaying graphs, classification results, and accuracy with the naïve bayes algorithm.
- Displays word cloud.
- Displays data sharing cross validation and accuracy.

b. Data Requirements Analysis

Analysis of Data Requirements to get the data needed in making this Sentiment Analysis System, the data needed are: Twitter user tweet data with keywords with hashtags used are Distance Learning System (DLS), Online Lectures (OL), and Online.

c. Environmental Needs Analysis

Analysis of the environmental requirements for the system is a first step to carry out a plan in the implementation of a Sentiment Analysis System for Distance Learning System (DLS). Environmental analysis was carried out to identify the need for DLS Sentiment information obtained from Twitter as the government's benchmark for the DLS method procurement. This is done using the help of the Analysis System in the form of a website. The website can be operated via a laptop or computer. Where, the author uses a laptop with Acer specifications (Swift SF314-5G), Intel® Core TM i7 2.0 GHz, 8GB RAM as a system creation tool. This laptop has the Windows 10 Operating System. The program used in making the Analysis System Namely; R programming for data analysis, RStudio for coding process with R programming language, and Rshiny for building websites interactive. Results Website can be displayed on the browser Microsoft Edge.

d. User Characteristics Analysis

Analysis of user characteristics is an identification of targets who will become users of a system that has been created. The user characteristics of the sentiment analysis system for distance learning are intended only for general users, who will later analyze sentiment data.

Data Collection Techniques

a. Data Source

Ddata used for this study use primary data. Primary data of the study were obtained from tweet data from Twitter users taken from the Twitter API.

b. Data Description

In this research data description is taken through tweet data from Twitter users via the Twitter API. From the data that has been obtained from the Twitter API, it consists of 600 text data with keywords using hashtags, namely; Analysis of the environmental requirements for the system is a first step to carry out a plan in the implementation of a Sentiment Analysis System for Distance Learning System (DLS), Online Lectures (OL), and online by manual crawling on twitter. After getting the data from crawling, the data will be grouped into 3 (three) classes to be labeled according to the type of sentiment which is a tweet with positive, neutral and negative sentiments by experts. Then, the results of labeling by experts are divided into two, namely training data and testing data, each with 480 training data and 120 testing data. This research uses sentiment classification with the Naïve Bayes method. Then, the test is done using Cross Validation to get the level of accuracy.

Research Stages

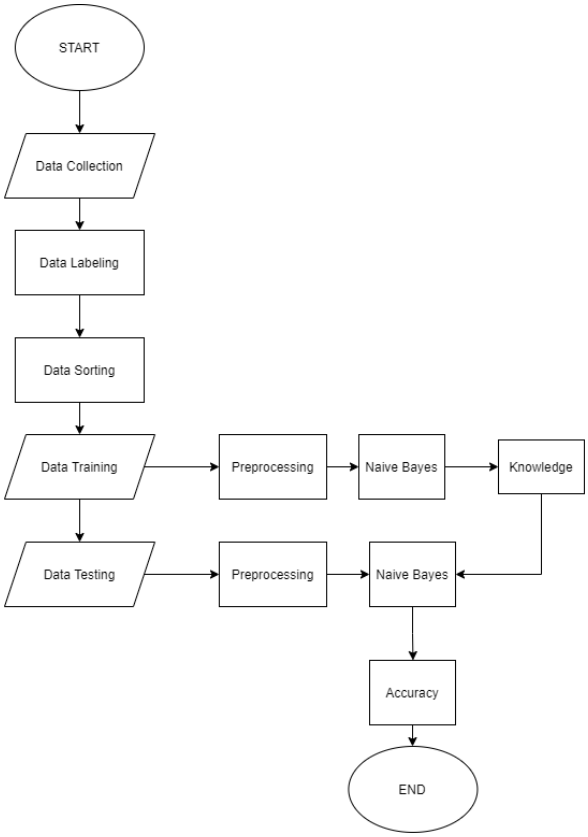

Figure 1. Research Flowchart

Figure 1 explains the stages of research on Sentiment Analysis on Distance Learning System (DLS) using the Naïve Bayes method. This stage begins with the collection of data taken from the use of hashtags on Twitter with keywords; Learning System (DLS), Online Lectures (OL), and online. The data obtained from Twitter is data from September to October 2020, where 600 sentiment data were collected. After the data is collected, the data will go

through the data labeling process by experts. The expert consists of 3 Lecturers at the Faculty of Economics and Business. Furthermore, the data will be processed through data sorting which is divided into 2 types, namely, training data and testing data. The training data will go through several processes including preprocessing. Preprocessing is one of the steps for selecting data to be processed. The preprocessing process carried out is Cleansing, Case Folding, Stopword Removal (Filtering), and Stemming. After getting the preprocessing results, the sentiment classification process will go through the naïve Bayes algorithm, resulting in labeling by the system in the form of positive, neutral and negative sentiment classification. After the training data, the sentiment classification can be used for testing data labeling. Meanwhile, testing data has the same process as training data. Then, the testing data and training data that have been labeled using the naïve Bayes method will produce accuracy.

Preprocessing Stages

Preprocessing is an early stage process for text mining change the unstructured information data so that it can make it easier in a research process according to the required format (Putra & Wardani, 2020). Process this is done to process and organize information by eliminating inappropriate information data so that it is easier for the system to process. As for preprocessing stages are divided into 4, including:

a. Cleansing

The cleansing process is carried out to clean the document from the said word not required. The omitted words, namely; HTML Characters, Emoticons, Hashtags, Username, URL (Uniform Resource Locator), Email, and "RT" (Retweet) (Rustiana & Rahayu, 2017).

b. Case Folding

The Case Folding process can change all letter characters in a sentence be lowercase, eliminating numbers, and punctuation marks (Rustiana & Rahayu, 2017).

c. Stopword Removal (Filtering)

The Stopword Removal process is a meaningless generic word important and not used. In this process unused words will be deleted to reduce the number of words stored by the system (Rustiana & Rahayu, 2017).

d. Stemming

Stemming process is the process of converting affixes into words base, by eliminating meaningless affixes (Rustiana & Rahayu, 2017).

Calculation Steps for Naïve Bayes

Naïve Bayes is an algorithm that has a classification method with uses a probability method based on the application of the Bayes Theorem where Classification is tried out through training sets of some information effectively. Naïve Bayes assumes the value of an input attribute in that class given independent of the value of other attributes (Dahri et al., 2016). The steps of calculating Naïve Bayes in this study are as following (Ruhyana, 2019):

a. The first step, calculating the conditional probability/ likelihood.

$$P(x|C) = P(x1,x2,\ldots,xn|C) \quad \ldots\ldots (1)$$

Description:
- C = Class.
- X = The vector of the attribute value n.
- P(xi|C) = Proportions of documents from class C.

b. The second step, calculating the prior probability for each class.

$$P(C) = \frac{Nj}{N} \quad \ldots\ldots (2)$$

Description:
- Nj = The number of documents in a class.
- N = Total number of documents.

c. The third step, calculating the posterior value.

$$P(C|X) = \frac{p(X|C).p(c)}{p(x)} \quad \ldots\ldots (3)$$

Description:
- X = Data with an unknown class.
- C = The data hypothesis is a specific class.
- P(c|x) = Hypothesis probability based on conditions (Posteriori Probability).
- P(c) = Hypothesis Probability (Prior Probability).
- P(x|c) = Probability based on the onditions in the hypothesis (likelihood).
- P(x) = Probability c.

d. The general formula for naïve bayes is as follows:

$$Posterior = \frac{likelihood\ x\ prior}{evidence} \quad \ldots\ldots (4)$$

Naïve Bayes Method Implementation Scenario

Implementing this Naïve Bayes Method using categories based on hypothetical Mean values and Hypothetical Deviation Standards with formula as follows (Maryam, 2018).

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v3i3.68

Table 1 Hypothetical Interval Distance Calculation Formula

Formula	Category
X < Mean – 1SD	Low
Mean – 1SD ≤ X < Mean + 1SD	Medium
Mean + 1SD ≥ X	High

Number of Items = 3
Xmin = 0
Xmax = 11
Range = 11
Mean = 6
SD = 2

Table 2 Category Calculation Results

Low	X < 6 – 1.2 X < 4
Medium	6 – 1.2 ≤ X < 6 + 1.2 6 – 2 ≤ X < 6 + 2 4 ≤ X 8
High	6 + 1.2 ≥ X 8 ≥ X

So, Low Values = 0 - 4, Medium Values = 5 - 8, and High Values = 9 – 12

a. P(C): Count the number of labels (Expert)
P(Negative) = 0,43
P(Neutral) = 0,4
P(Positive) = 0,18

b. P(D|C): Based on negative words
P(Low|Negative) = 0,63
P(Medium|Negative) = 0,25
P(High|Negative) = 0,12

c. P(D|C): Based on the word neutral
P(Low|Neutral) = 0,63
P(Medium|Neutral) = 0,23
P(High|Neutral) = 0,15

d. P(D|C): Based on Positive Words
P(Low|Positive) = 0,48
P(Medium|Positive) = 0,33
P(High|Positive) = 0,19

Negative = 0.008127
Neutral = 0.008694
Positive = 0.005417

then, the greatest is the probability value for the class Neutral with a Value of 0.008694.

RESULTS AND DISCUSSION

System Results

The sentiment analysis system from distance learning data research is as follows:

a. Displays Training Data

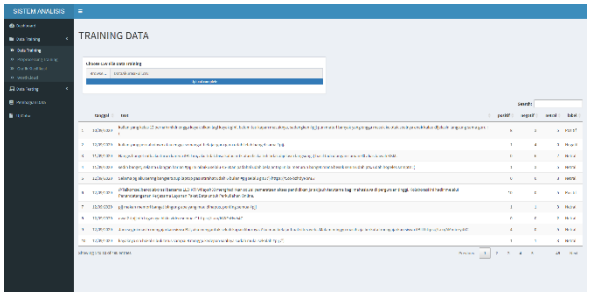

Figure 2. Enter training data

Figure 2 is the process of entering training data by the user and displaying the data.

b. Displays Preprocessing Training

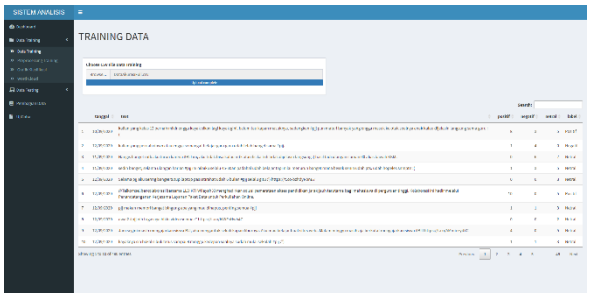

Figure 3. Preprocessing Training

Figure 3 is a process that displays the results of preprocessing training data.

c. Graph Displays of Training Classification

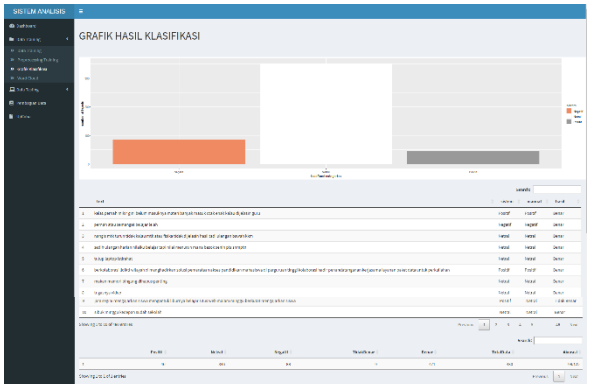

Figure 4. Training Classification Graph

Figure 4 is a Graph system, Classification Results, and 80% accuracy of training data with the Naïve Bayes method resulting in 46 positive, 346 neutral, and 88 negative. So, the public sentiment towards the Implementation of Distance Learning System (DLS) is Neutral with an accuracy of 98%.

d. Displays Word Cloud Training

Figure 5. Word Cloud Training

Figure 5 is the process that displays the Word Cloud on the Training Data with the words that come out the most are learning, assignments, teachers, quotas, and schools.

e. Displays Test Data

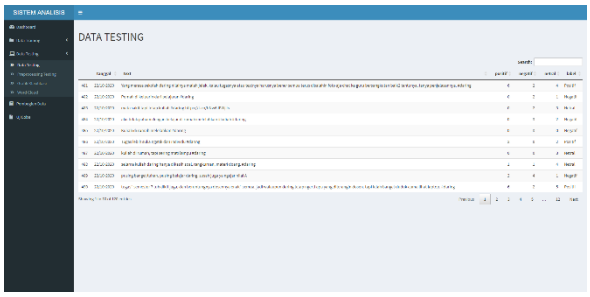

Figure 6. Test Data

Figure 6 is the process of displaying test data.

f. Displays Testing Preprocessing

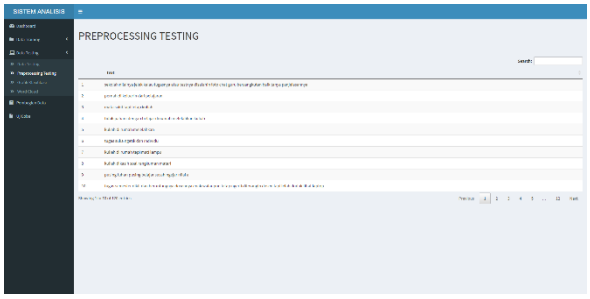

Figure 7. Testing Preprocessing

Figure 7 shows the results of the data testing preprocessing process.

g. Graph Displays of Testing Classification

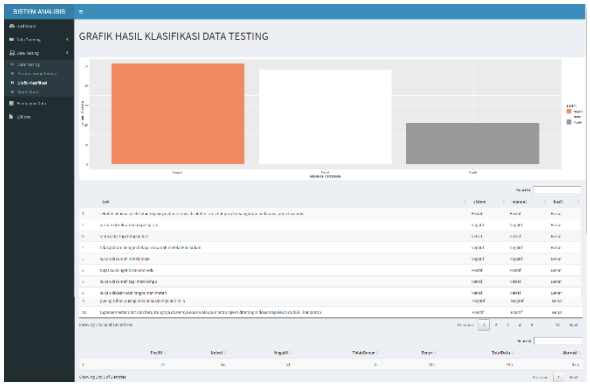

Figure 8. Testing Classification Graph

Figure 8 is a graph system, classification results, and 20% accuracy. Data testing with the Naïve Bayes method resulted in 21 positive, 48 neutral, and 51 negative. So the public sentiment towards the Implementation of Distance Learning System (DLS) is negative with an accuracy of 98%.

h. Displays Word Cloud Testing

Figure 9. Word Cloud Testing

Figure 9 is a process that displays Word Cloud on Data Testing with the most common words being study, school, and home.

i. Displays Data Sharing

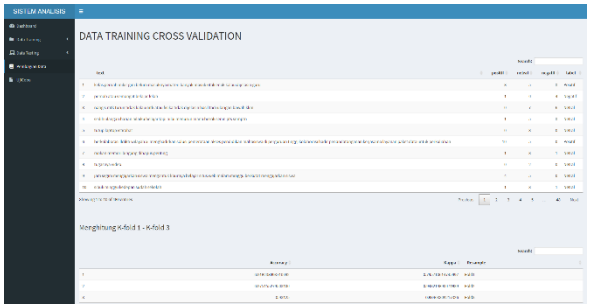

Figure 10. Data Sharing

Figure 10 is a process that displays the 3-fold distribution of Cross Validation data with 80%

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v3i3.68

training data and 20% testing data. With an average training data accuracy of 94%. While the average test data is 93%.

j. Displays Trial

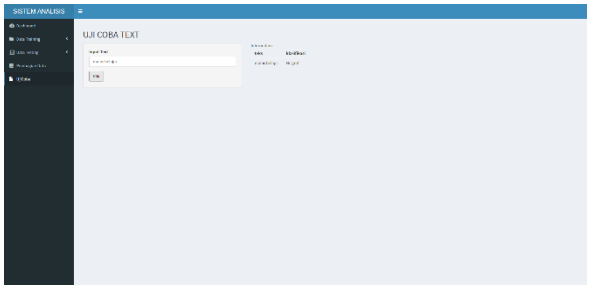

Figure 11. Trial

Figure 11 shows the text testing process to be classified using the naïve Bayes method.

Evaluation Result

Table 1. Naïve Bayes Scenario with 3-fold Cross Validation Accuracy

Training		Testing	
Data K-	Accuracy	Data K-	Accuracy
1	94%	1	90%
2	94%	2	92%
3	95%	3	97%
Average	94%	Avarage	93%

The table above is a naïve Bayes scenario with 3 times cross validation accuracy using 600 sentiment data. The sentiment data used is about the implementation of Distance Learning System (DLS) which is taken from Twitter data using the Twitter API from September to October 2020. The data is divided into 80% Training Data and 20% Testing Data. The training data is used to train the algorithm, while the testing data is used to determine the performance of the algorithm that has been trained. The average accuracy of Cross Validation 3 times results in 94% of Data Training and 93% of Data Testing.

CONCLUSIONS AND SUGGESTIONS

Conclusion

After conducting research on Distance Learning System (DLS) sentiment analysis using Naïve Bayes Algorithm, it can be concluded that, Public opinion on Distance Learning Sentiment Data in Indonesia from September to October generates negative sentiment due to many obstacles during the learning process conducted at home. From the data classification results, public sentiment analysis of DLS using Naïve Bayes method is very good with the accuracy of sentiment analysis on distance learning using Naive Bayes method using 3 fold cross validation accuracy resulting in an average of 93%.

Suggestion

It is hoped that the distance learning process can better suit the needs of the community and the system can be developed even better.

REFERENCES

Ahmad, I. F. (2020). Asesmen Alternatif Dalam Pembelajaran Jarak Jauh Pada Masa Darurat Penyebaran Coronavirus Disease (Covid-19) Di Indonesia. *PEDAGOGIK: Jurnal Pendidikan*, *7*(1), 195–222. https://doi.org/10.33650/pjp.v7i1.1136

CAKTI INDRA GUNAWAN, S. E. M. M., & YULITA, S. E. M. A. P. (2020). *ANOMALI COVID-19 : DAMPAK POSITIF VIRUS CORONA UNTUK DUNIA*. IRDH Book Publisher. https://books.google.co.id/books?id=CWzuDwAAQBAJ

Dahri, D., Agus, F., & Khairina, D. M. (2016). Metode Naive Bayes Untuk Penentuan Penerima Beasiswa Bidikmisi Universitas Mulawarman. *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, *11*(2), 29. https://doi.org/10.30872/jim.v11i2.211

Devita, R. N., Herwanto, H. W., & Wibawa, A. P. (2018). Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(4), 427.https://doi.org/10.25126/jtiik.201854773

Maryam, E. W. (2018). Gambaran Sense Of Community Pada Karyawan Bagian Administrasi Di Universitas Muhammadiyah Sidoarjo. *Psikologia : Jurnal Psikologi*, *2*(1), 52.https://doi.org/10.21070/psikologia.v2i1.756

Praningki, T., & Budi, I. (2018). Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN. *Creative Information Technology Journal*, *4*(2), 83. https://doi.org/10.24076/citec.2017v4i2.100

Putra, M. P. R., & Wardani, K. R. N. (2020). Penerapan Text Mining Dalam Menganalisis Kepribadian Pengguna Media Sosial. *JUTIM (Jurnal Teknik Informatika Musirawas)*, *5*(1),

63–71. https://doi.org/10.32767/jutim.v5i1.791

Rifqo, M. H., & Wijaya, A. (2017). Implementasi Algoritma Naive Bayes Dalam Penentuan Pemberian Kredit. *Pseudocode*, *4*(2), 120–128. https://doi.org/10.33369/pseudocode.4.2.120-128

Risnantoyo, R., Nugroho, A., & Mandara, K. (2020). Sentiment Analysis on Corona Virus Pandemic Using Machine Learning Algorithm. *Journal of Informatics and Telecommunication Engineering*, *4*(1), 86–96. https://doi.org/10.31289/jite.v4i1.3798

Rosyad, N. N. (2019). *Analisis Sentimen Publik Terhadap Sistem Zonasi Sekolah Menggunakan Data Twitter Dengan Metode Naïve Bayes Classification*. *12*(4), 315–322. https://doi.org/10.30998/faktorexacta.v12i4.5205

Ruhyana, N. (2019). Analisis Sentimen Terhadap Penerapan Sistem Plat Nomor Ganjil / Genap Pada Twitter Dengan Metode Klasifikasi Naive Bayes. *Jurnal IKRA-ITH Informatika*, *3*(1), 94–99.

Rustiana, D., & Rahayu, N. (2017). Analisis sentimen pasar otomotif mobil: *Jurnal SIMETRIS*, *8*(1), 113–120.

Salam, M. A. K. (2020). Perilaku Produksi di Tengah Krisis Global Akibat Pandemi Covid-19 dan Memanfaatkan Media Online Facebook Sebagai Alternatif Pasar. *Ekonomi, Manajemen Dan Akuntansi ISSN: 1979-9888*, 1–21. http://eprints.umsida.ac.id/id/eprint/6834

Siagian, T. H. (2020). Mencari Kelompok Beresiko Tinggi Terinfeksi Virus Corona Dengan Discourse Network Analysis. *Jurnal Kebijakan Kesehatan Indonesia*, *09*(02), 98.

Sulaiman, O. K. (2020). *Merdeka Kreatif di Era Pandemi Covid-19* (Issue August).