ISSN No:-2456-2165

Online School Sentiment Analysis in Indonesia on Twitter Using The Naïve Bayes Classifier and Rapid Miner Tools

Ahmad Cahyono Adi, Dyan Puji Lestari, Elsa, Fiqrudina Sain Saputri, Yohanes Sabui
Faculty of Mathematics and Natural Sciences
Tanjungpura University
Indonesia

Abstract:- The COVID-19 pandemic entered at the beginning of 2020 which hit various countries in the world, including Indonesia, with 4,259,644 contaminated cases (Kawal Covid 19. 2021). The impact of the COVID-19 pandemic is in the economic, tourism, and education sectors. The most obvious impact due to this pandemic is in the field of education, where every process of teaching and learning activities is limited or even encouraged to study from home. Therefore, teaching and learning activities are carried out online or in a network (online). Educators are starting to look for alternative methods used in online learning because in Indonesia they still use conventional learning or are still in the form of face-to-face learning directly with a classical system.The Naive Bayes method is a classification using probability and statistical methods, namely by predicting future opportunities based on previous experience. The main feature of the Naïve Bayes classification is to get a strong hypothesis from each condition or event. The following is the equation of Bayes' theorem In the rapid miner tools, the data is then retrieved by taking the text and label attributes that have been given during the labeling process. The data is added with a label attribute to facilitate the classification process. After that, the labeled data is then normalized by changing or removing unimportant attributes, in this case, the unimportant data is other than letters. Therefore, a process of deleting non-letter-type data is needed which includes numbers and symbols.

Keywords:- *Covid-19; Naïve Bayes Algorithm;Sentiment Analysis; Text Mining;Twitter.*

I. INTRODUCTION

The COVID-19 pandemic entered at the beginning of 2020 which hit various countries in the world, including Indonesia, with 4,259,644 contaminated cases [1].The impact of the COVID-19 pandemic is in the economic, tourism, and education sectors. The most obvious impact due to this pandemic is in the field of education, where every process of teaching and learning activities is limited or even encouraged to study from home. Therefore, teaching and learning activities are carried out online or in a network (online). Educators are starting to look for alternative methods used in online learning because in Indonesia they still use conventional learning or are still in the form of face-to-face learning directly with a classical system. In the past year, in Indonesia, there has been a lot of online learning in schools and even campuses in Indonesia. As a simple step, online learning begins by utilizing existing social media, such as Whatsapp, telegram, and youtube. In addition, applications used for the learning process such as Google Classroom, Edmodo, Google Meet, or Zoom [2].

The existence of online schools has attracted a lot of responses from various levels of society, both agree and disagree. Complaints to expressions of pleasure are found on various social media platforms. In addition to complaints and expressions of pleasure, not a few also provide suggestions to improve the education system in online conditions [3].

One of the most widely used social media as a medium for expressing opinions and exchanging ideas is Twitter. Likewise, during the pandemic, Twitter is still one of the social media that has high popularity when compared to other social media. This is evidenced by data from the Ministry of Communication and Information (2021) which states that Indonesia is ranked 6th with 15.7 million Twitter users. The activities of Twitter users who are interested in being active in conversations and dominated by interesting topics lead to interaction so that freedom of expression is built on this social media [4]. Due to a large number of opinions regarding online schools during the Covid-19 pandemic, it is necessary to analyze sentiment on this phenomenon as a consideration so that related parties can consider it in order to overcome the problems that occur. Sentiment analysis is one of the analytical methods from Text Mining that can be used to classify documents in the form of opinion texts based on sentiment [5].Sentiment analysis aims to determine attitudes towards several topics as well as the contextual polarity of the entire document [3].

II. MATERIALS AND METHODS

This research method uses the Naïve Bayes method in classifying data to obtain results. The following are the stages of the research described in the image below.

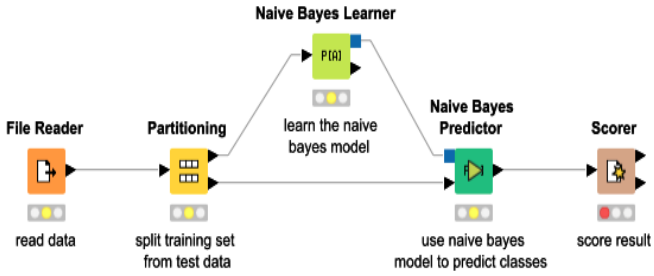

Fig. 1: Naïve Bayes Learner

ISSN No:-2456-2165

The picture above shows an opinion mining technique through Twitter using the Naïve Bayes method. Crawling data is carried out by providing keywords within a certain period. Furthermore, the data will be normalized using preprocessing data. The data preprocessing stage aims to select the data and change it to make it more structured. At this stage, the data cleaning process is carried out to reduce the possibility of noise data and remove stop words[6].Furthermore, the tokenization process is used to identify words and break sentences into terms based on spaces and punctuation marks [7]. Furthermore, the last stage in preprocessing is stemming, changing affixes into basic words [8]. The third stage in opinion mining is to perform feature extraction to simplify the Naïve Bayes classification. This stage generates a model and is used to show the accuracy of the classification results[9].

The data in tweet format is crawled from Twitter social media and saved in Excel file format. The data were divided into two data sets: training data and test data. There are labels to distinguish positive, neutral, and negative tweets. The naive Bayes method is used at the stage of classification and interpretation of the results of sentiment analysis.

The Naive Bayes method is a classification using probability and statistical methods, namely by predicting future opportunities based on previous experience [10].The main feature of the Naïve Bayes classification is to get a strong hypothesis from each condition or event [11].The following is the equation of Bayes' theorem [12].

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

Where:
X : Data with unknown class
H : Hypothesis data is a specific class
P(H|X): Probability of Hypothesis H based on condition X (posterior probability)
P (H): Hypothesis probability H (prior probability)
P(X|H) : probability X based on condition on hypothesis H
P(X) : Probability X

Naive Bayes is based on the simplifying assumption that attribute values are conditionally independent if given an output value. In other words, given the output value, the probability of observing together is the product of the individual probabilities [13]. The advantage of using Naive Bayes is that this method only requires a small amount of training data (Training Data) to determine the parameter estimates needed in the classification process. Naive Bayes often performs much better in most complex real-world situations than one might expect[14].Naïve Bayes algoritm have a better accuracy than any algoritm, because this algoritm use probability and statistics method[15].

III. RESULT

	true netral	true negatif	true positif	class precision
pred. netral	73	11	25	66.97%
pred. negatif	10	19	11	47.50%
pred. positif	26	10	27	42.86%
class recall	66.97%	47.50%	42.86%	

Fig.3: Result

The results of the classification process can be seen in the figure 3, where based on the confusion matrix that has been shown, each precision class has a different percentage with 66.97% for the neutral precision class, 42.86% for the positive precision class, and 47.50% for the negative precision class with the resulting accuracy rate is 42.86% for positive, 47.50% for negative and 66.97% for neutral.

In rapid miner, the naive Bayes algorithm has been provided so that its use only needs to adjust the parameters so that the data can be processed. After that, the results of the classification using the naive Bayes algorithm are visualized to see the level of accuracy produced. Visualize the data using a confusion matrix to see the precision and accuracy produced.

IV. DISCUSSION

In general, this research method can be divided into data collection, data pre-processing, data analysis, and data visualization. In this study, data was taken using the export comment tool, the data needed was YouTube commentary data related to the online School Tweet in Twitter.

The data is taken using the Twitter link and will be converted into excel-based data. In this study, 223 comments were obtained which will be analyzed based on sentiment to find out and classify netizen responses. Data preprocessing is an early data mining technique to convert raw data or commonly known as raw data collected from various sources into cleaner information that can be used for further processing. This process can also be called the initial step to retrieve all available information by cleaning, filtering, and combining data. 3 common problems solved in the preprocessing stage are dealing with missing values, noise data, and inconsistent data. Missing values are inaccurate data because missing information makes the information in them irrelevant. Missing value often occurs when there is a problem with the collection process, such as an error in data entry or a problem with the use of biometrics. Noise data contains erroneous data and outliers that can be found in the data set. These outliers and erroneous data contain meaningless information. Some of the causes of data noise are due to human errors in the form of labeling errors and other problems during data collection[16].

Comment (Text)	Label
@tubirfess online aja nggak semangat.. apalgi offline yg dijamin bakal lebih berat/hectic.. disamping pengeluaran yg lebih banyak juga	Negative
[CM] kok bisa yaa temen2 di menfess ini rajin bgt nyatet selama kuliah online. aku malah semester ini gak nyatet sama sekali, kalo kuliah offline baru aku rajin nyatet😊	Negative
@collegemenfess Aku tim buat catatan sebelum kuliah 😊Jadi mmyg mau ditanyakanpun udah dilist hehe	Positive
@collegemenfess Kok kita sama nder	Netral

Table Data Normalization

Preprocessing data is very important because errors, redundancy, missing values, and inconsistent data lead to reduced accuracy of analysis results. In this research, The data that has been taken is then carried out a normalization process to eliminate some items that are not important so that the resulting data does not contain noise that can interfere with the classification process. Preprocessing data is a process to cleaning and make sure that the data is vali. Data preprocessing use for avoid the failed, inconsistent, or duplicate data [17].In the first preprocessing data, the comment data that has been taken is then given a label that includes positive, negative, and neutral labels. The label is assigned manually. Data that has been labeled will then be processed using a rapid miner. In the rapid miner tools, the data is then retrieved by taking the text and label attributes that have been given during the labeling process. The data is added with a label attribute to facilitate the classification process. After that, the labeled data is then normalized by changing or removing unimportant attributes, in this case, the unimportant data is other than letters. Therefore, a process of deleting non-letter-type data is needed which includes numbers and symbols.

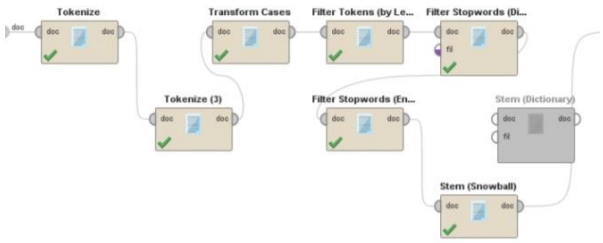

Fig 2. Data Analysis Rapid Miner
(Data Analysis)

After the data has been normalized by deleting non-letter type data, then the data will be tokenized. In this study, the tokenize used is 2. The first tokenize is to remove symbols, the second tokenize is to delete non-letter data. Furthermore, the data is converted into lowercase data to facilitate the classification process, the data is then filtered to remove unimportant letters, using a filter by a length with a minimum letter length parameter of 2 and a maximum of 25 with the aim of letters or words that are outside from that range will not be processed. A process is carried out to filter the stopword, the stopword filter used is the stopword filter in Indonesian, and the stopword filter in English, the goal is to remove the affixed words to become a verb that can be analyzed.the data preprocessing is complete, it is continued with data analysis using the naive Bayes algorithm.

REFERENCES

[1.] Anonymous.2020.Kawal Covid-19 https://kawalcovid19.id/ Access on 6 January 2022
[2.] Kutsiyyah.2021.”Analisis Fenomena Pembelajaran Daring Pada Masa Pandemi (Harapan Menuju Blended Learning)”.Edukatif : Jurnal Ilmu Pendidikan Vol.3 No.4 pp:1460-1469.
[3.] Ratnawati Fajar.2018.” Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter”. JURNAL INOVTEK POLBENG - SERI INFORMATIKA, VOL. 3, NO. 1.50-59.
[4.] Savitri C.P. Luh Ni et al.2020.”Analisis Klasifikasi Sentimen Terhadap Sekolah Daring pada Twitter Menggunakan Supervised Machine Learning”. Jurnal Teknik Informatikadan Sistem Informasi Vol.7 No.7 pp:47-58.
[5.] Juditha Christiany.2018.”Interaksi Komunikasi Hoax di Media Sosialserta Antisipasinya Hoax Communication Interactivity in Social Media and Anticipation”. Jurnal Pekommas, Vol. 3 No. 1 pp:31-44.
[6.] Rofiqoh Ummiet al.2017.”Analisis Sentimen TingkatKepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexicon Based Features”.Jurnal Pengembangan Teknologi Informasidan Ilmu Komputer Vol.1, No.12 pp:1725-1732.
[7.] R. Watrianthos, 2019.“Analisis Pembelajaran Daring di Era Pandemic Covid-19,” in MerdekaKreatif di Era Pandemi Covid19: SuatuPengantar, Medan: Green Press, 2020, p. 55.
[8.] A. Sholihin, Haviluddin, N. Puspitasari, M. Wati, and Islamiyah, 2019. “Analisis Penyakit Difteri Berbasis Twitter Menggunakan Algoritma Naïve Bayes,” SAKTI – Sains, Apl. Komputasidan Teknol. Inf., vol. 1, no. 1, pp. 7–15,
[9.] Eden Billy dkk. 2018.” Sistem Informasi Peramalan Harga Pangan Dengan Menggunakan Metode Naïve Bayes Di Kota Makassar”. JURNAL SISTEM INFORMASI DAN TEKNOLOGI INFORMASI Vol. 7, No. 2.163-171.
[10.] D. Setian and I. Seprina.2020. “ANALISIS SENTIMEN MASYARAKAT TERHADAP DATA TWEET LAZADA MENGGUNAKAN TEXT MINING DAN ALGORITMA NAIVE BAYES CLASSIFIER,” in Bina Darma Conference on Computer Science.pp. 998–1004.
[11.] R. Watrianthos, S. Suryadi.D. Irmayani, M. Nasution, and E. F. S. Simanjorang. 2019.“Sentiment Analysis Of Traveloka App Using Naïve Bayes Classifier Method,” Int. J. Sci. Technol. Res., vol. 8, no. 07, pp. 786–788.
[12.] N. Rochmawati and S. C. Wibawa.2018 “Opinion Analysis on Rohingya using Twitter Data,”.IOP Conf. Ser. Mater. Sci. Eng., vol. 336, no. 1, ,doi: 10.1088/1757-899X/336/1/012013.

[13.] Saputra, R. A., Taufik, A. R., Ramdhani, L. S., Oktapian, R., &Marsusanti, E. (2018). SistemPendukungKeputusanDalamMenentukanMetode KontrasepsiMenggunakanAlgoritma Naive Bayes. SNIT 2018, 106–111.
[14.] Ridwan, M., Suyono, H., Sarosa, M., 2013, Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier, Jurnal EECCIS, Vol 1, No. 7, Hal. 59-64
[15.] Bustami., 2013, Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, TECHSI :Jurnal Penelitian Teknik Informatika, Vol. 3, No.2, Hal. 127-146.
[16.] Mustafa, M.S., Ramadhan, M.R., and Thenata, A.P. (2018).Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier. Citec Journal. Vol. 4(2): 151-162.
[17.] Oliver Andrea.(2021).”Bikin data lebihmudahdibaca, yuk kenalandengan data preprocessing”.Glint.com https://glints.com/id/lowongan/data-preprocessing-adalah/#.YeLGWugzbIV access on 5 January 2022
[18.] Asroni, Fitri, H., and Prasetyo, E. 2018. Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (StudiKasus: Fakultas Kedokterandan Ilmu Kesehatan, dan Fakultas Ilmu Sosialdan Ilmu Politik. Jurnal Semesta Teknika. Vol. 21(1): 60-64.