SENTIMENT ANALYSIS ON INVESTMENT TOPIC IN INDONESIA USING MACHINE LEARNING ALGORITHMS APPROACH

FLAVEGA MONESA[1], RIYANTO JAYADI[2]

[1,2] Information Systems Management Department,
Binus Graduate Program – Master of Information Systems Management,
Bina Nusantara University
E-mail: [1]flavega.monesa@binus.ac.id, [2]riyanto.jayadi@binus.edu

ABSTRACT

Investment is one of the rising topics in Indonesia after the pandemic. To be able to understand the overall sentiment or public opinion towards the investment topic, it is important to monitor its growth of it through opinions on social media. In this study, an automated sentiment analysis on the Investment topic in Indonesia is conducted. The proposed technique automatically labeled the sentiment as positive and negative. The opinion data used is in Bahasa Indonesia. To train the model, opinions were scrapped from tweets through Twitter Developer API. There are 10,000 tweets in Indonesian scrapped from September to November 2022 using "Investasi" keyword This analysis uses Naïve Bayes Classifier and Support Vector Machine model to compare which algorithms has better accuracy. The result shows that the SVM algorithm works better with an accuracy rate of 95,7% compare to Naïve Bayes Classifier with an accuracy rate of 94,6%.

Keywords: *NLP, Sentiment Analysis, Machine Learning, Naïve Bayes, Support Vector Machine (SVM)*

1. INTRODUCTION

Investment in understanding is something of value that is purchased to generate profits or make more money. Investments are usually associated with stocks, bonds, and other financial instruments but investments can also take the form of real-estate, works of art, collectibles, or even wine. Talking about investments is usually always followed by risks. Whatever form of investment you choose, there will always be risks in every instrument you choose [1]. In Indonesia, there are various types of investment options available, but of course, before investing there are various factors that investors need to pay attention to in choosing investment instruments, namely investors need to pay attention to investment objectives, risk profiles, and also the timeframe for investing where the investment period is divided into two, namely investment short term and long-term investment. For short-term investments, usually, the results of the investment can be seen in a period of 3 to 12 months where usually for short-term investments the level of liquidity tends to be high or can be easily resold, and usually, short-term investments have relatively low returns, one of the most common forms of investment. Popular for the short term are mutual funds. Long-term investments usually take years to get returns, and it is not uncommon for someone to be able to enjoy investment returns after dozens of years, but in contrast to short-term investments, long-term investments tend to provide quite large returns and the risks borne also tend to be greater. usually in the first year, the form of long-term investment gives losses instead of profits, one example of a form of long-term investment that is quite popular, namely stocks, as for the

forms of investment, namely Deposits, Gold, Property, Stocks, Mutual Funds and Peer to peer lending [2] Even though the Covid-19 pandemic has paralyzed the world economy globally, in fact, the Indonesian people are starting to be literate in investment or starting to have an awareness of the importance of investment from an early age and during the pandemic, retail investors in Indonesia began to be busy investing in mutual funds and stocks in line with the literacy of the Indonesian people with investments starting you can find many investment platforms that educate the public about how to invest and choose the right instruments. A study conducted by RSCH8083-Thesis (Colloquium) Investor Global Schroders 2021 found that almost a third of investors globally invested more during the Covid-19 pandemic, this condition caused investors to think more about their financial well-being, and this trend is also expected to continue, while the researcher surveyed around 23,000 people from 32 locations globally, and found that almost half of the investors, namely 46% of those aged 18-37 years, will now save their funds to invest more after social restrictions are lifted by the government and this, of course, includes Indonesia [3]. The increase in retail investors in Indonesia is dominated by the Millennial generation, this is quoting from the Faculty of Economics and Business, Gadjah Mada University (KAFEGAMA) explained in Figure 1.1 that in 2021 discussing opportunities, challenges, and policies in line with the increase in retail investors as for Inardo's explanation Djajadi said that the total investors as of September 2021 were 6,431, stock investors were 2.9 million investors and there was an increase of 7.2% from stock investors which

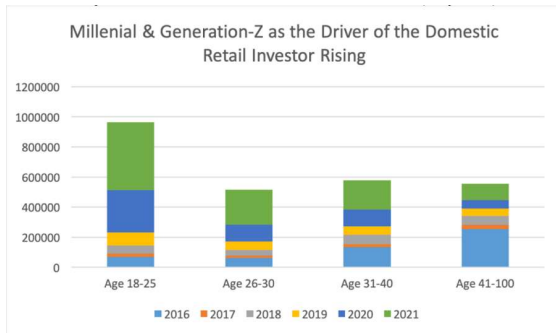according to him this achievement could not be separated from the influence of social media influencers[4].

Figure 1. Increase in Retail Investors Based on Age in Indonesia [4]

With the spread of Indonesian public awareness of the importance of investment, various platforms can disseminate information quickly, one of which is social media where social media is the platform used and facilitates the dissemination of information more quickly and efficiently. According to a report from We Are Social, as shown in Figure 1.2, it is explained that Indonesia is a country whose people are quite active in using social media where there are 191 million active social media users in Indonesia in January 2022, this figure represents an increase from last year's 170 million people, an increase of 12 percent 35% then on the trend Indonesia continues to experience growth in the number of active social media users from 2014 to 2022.
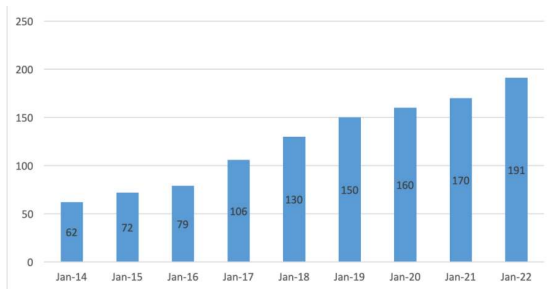

Figure 2. Number of Active Social Media Users [5]

For social media that is popularly used in Indonesia, one of them is Twitter which is a concept of social networking and microblogging. Twitter itself in Indonesia has quite a lot of users. Twitter has unique features such as direct messages, follows followers, tweets, and retweets where tweets or tweets are content that has a character limit so that discussions on Twitter often tend to be to-the-point and concise. Based on statistical data, it can be seen in Figure 1.3 that Twitter has 206 million daily active users worldwide in the second quarter of 2021 where the United States is the country with the most Twitter users in the world, while Indonesia is ranked 6th with the most Twitter users in the world, namely by the

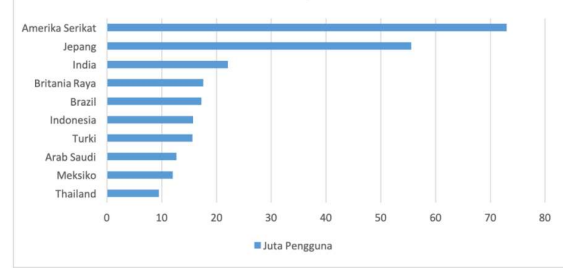number users by 15.7 million in 2021[6]

Figure 3. Countries with the most Twitter User [6].

This sentiment analysis is expected to be used to determine the sentiment of the Indonesian people toward investment. Sentiment analysis is a key instrument in social media marketing strategies where by analyzing customer sentiment about a topic, product or service. Companies can determine strategies that can be carried out for the benefit of the company. Sentiment analysis is a tool that automatically monitors the emotions contained in conversations on social media platforms, one of which is tweets or tweets on Twitter. With sentiment analysis data on investment in Indonesia, it is hoped that it can become a reference for investment companies to develop forms or investment products that are in line with the sentiments of the Indonesian people and naturally, this sentiment analysis is also expected to be able to help the Indonesian people to capture investment trends in Indonesia. In previous research, research has been conducted on public opinion sentiment on Twitter social media regarding investment applications in Indonesia where this research discusses investment applications in Indonesia and aims to analyze which investment applications need to be avoided and which can be trusted. This study uses the Naïve Bayes method and the lexicon-based to classify tweets or tweets with positive, negative, or neutral sentiments to make it easier for people to make choices and compare the level of accuracy between the two methods [7] As for other research that also discusses investment in Indonesia, more precisely on the ratification of the Omnibus Law Bill in the Investment and Employment sector where this research uses the Naïve Bayes in analyzing public perceptions using data from Twitter tweets in 2020 and this research finds that some of the words in tweets discuss employment and based on analysis the word "investment" has a close relationship with Indonesia's economic growth while the word copyright is closely related to job creation and the word labor has a close relationship with rejection and then the results of the sentiment analysis found 54% of tweets were classified as negative sentiment 46% of tweets classified as positive sentiment and then using the Naïve Bayes method the model formed can

classify 87.1% of Twitter tweets correctly or an accuracy rate of 87.1% [8]. Then there are also previous studies that conducted research on sentiment analysis on investment applications in Indonesia, namely bibit and bareksa where this study aims to analyze sentiments on user reviews of bibit and bareksa investment applications with a total of 998 reviews consisting of 484 positive sentiments and 524 negative sentiments for the bareksa application and 1063 reviews consisting of 542 positive sentiments and 522 negative sentiments and this data was processed using the CRISP-DM with the algorithm used, namely K-Nearest Neighbor with a model accuracy of 85.4% for seedlings and 81.70% for bareksa [9]. In writing this research, it is hoped that the writer will be able to examine information or trends contained in the topic of investment in Indonesia and be able to read an overview of the perspectives or sentiments of the Indonesian people regarding investment. The keywords that will be used in data collection are "Investment", "shares", "bonds", "mutual funds", and "deposits" to be able to obtain data related to investment in Indonesia and specifically considering stocks, bonds, mutual funds, and Deposits are a popular form of investment in Indonesia. Writing uses the Naïve Bayes classifier method, Support Vector Machine (SVM), Decision Tree, Random Forest, Neural Network, and Gradient boost in seeing the sentiment of Indonesian public opinion on investment topics and it is hoped that this research can be a reference or for companies that provide investment platforms to issue strategies that can help the development of investment products as well as become a reference for Indonesian people who want to invest with this sentiment analysis.

2. LITERATURE REVIEW

2.1 Investment

Investment in economics can mean investment made by local investors, commonly referred to as a domestic investment, then foreign investors are called foreign investment, and investment is made indirectly by foreign parties. The narrative from Ida Bagus Rahmadi Supanca defines investment as an activity carried out by a person or legal entity to increase or maintain the value of their capital, whether in the form of cash, equipment, immovable assets, intellectual property rights, or expertise [10].

2.2 Social Media

In the technological age that we live in today, humans socialize more quickly and effectively with

social media in which social media is a platform that allows humans or communities to share information quickly and efficiently whether it's sending messages, looking for hot topics, or looking for goods or information. make social media needed by humans. It can be said that social media is a digital platform that provides socialization facilities to its users whether it's exchanging information or content which can be in the form of writing, video, audio, and photos. The following is an explanation from several experts regarding social media, the first from B.K Lewis through his book entitled Social Media and Strategic Communication Attitudes and Perceptions among College Students, which was published in 2010. He explained that social media is a label that refers to digital technology that has the potential to everyone connected and interacting, producing and sharing messages. Then the narrative from Chris Bogan from the same year in his book entitled Social Media 101: Tactics and Tips to Develop your Business explains that social media is a communication tool that contains various possibilities to create new forms of interaction. Main objectives of this research are the following ;

✓ To Predict a sentiment of a Tweet with a pre-determined keyword classified in two class which are positive and negative.
✓ To Understand Public Sentiment towards Investment topic in Indonesia as a whole.
✓ To give business insights to be used as a reference to develop investment in Indonesia with a Word-cloud.

2.2.1 Twitter

Twitter is a social media platform that allows its users to socialize. There are interesting features that make it easier for Twitter users to use the Tweet or tweet feature in uploading photos, videos, links, audio and even text [11].
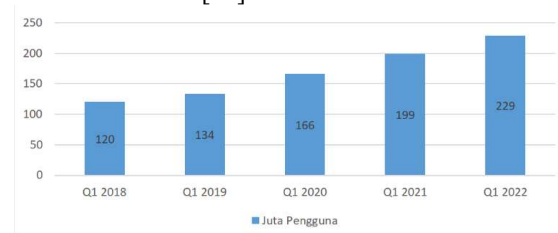

Figure 4 Number of Daily Monetizable Twitter Active Users (mDAU)[6].

The number of Twitter users continues to climb from year to year, as can be seen from figure 4 from Q1 2018 to Q1 2022, the number of monetizable daily active users (mDAU) in 2018 was initially only

around 120 million people, then the number continued increased to 199 million people in QI 2021 then for the Q1 2022 period the active users changed to 229 million people[12]. Data from We Are Social as can be seen in Figure 4 that Twitter is included as one of the ten most used social media in Indonesia where Twitter ranks sixth with a percentage of 58.3% so this will make it easier for researchers to obtain data for this project, namely to find out the sentiments of the Indonesian people regarding investment.

2.3 Sentiment Analysis

Sentiment analysis or commonly referred to as Sentiment Analysis is the process of using Text Mining or text mining that identifies and pulls subjective information from material sources and is used to help businesses understand or find out social sentiment towards the brands, products, or services they offer by monitoring online conversations. [13]. Sentiment analysis is included in the field of Natural Language Processing (NLP) to build a system to be able to recognize and draw an opinion from a text.

2.4 Machine Learning
Machine Learning is a branch of artificial intelligence or what is commonly called Artificial Intelligence wherein AI is generally divided into seven branches, namely Machine Learning, Natural Language Processing, expert systems, vision, speech, planning, and robotics. Machine Learning (ML) technology is developed so that machines can learn on their own without help from the developer. The term ML was first put forward by a mathematician named Adrien Marie Legendre, Thomas Bayes, and Andrey Markov in 1920. There are two basic techniques for learning from ML, namely:

1. Supervised Learning
In this technique, the user trains the machine using well-labeled data, which means that some data has been tagged with the correct answer. This algorithm learns from labeled training data, helping users predict the results of unexpected data.

2. Unsupervised Learning

Unsupervised Learning is an ML technique where the user does not need to supervise the model but rather lets the machine work on its own to find information, unlike Supervised data. Unsupervised learning is not labeled [14].

2.4.1 Naïve Bayes Classifier

Classification has the meaning of a machine learning model that is used to distinguish different objects based on certain features. Then principle

Naïve Bayes is a probabilistic machine learning model used for classification tasks and has a core sourced from Bayes' theorem[15]. The following is the form of the formula for the Bayes theorem as shown in Figure 5:

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

Figure 5 Bayes Theorem [15].

2.4.2 Support Vector Machine (SVM)
Support Vector Machine is a set of Supervised Machine Learning methods commonly used for classification, regression and for outlier detection. The advantage of using SVM is its effectiveness in high-dimensional space and it is also effective in cases where the dimensions are larger than the number of samples and it uses a subset of decision function training points or what are commonly called support vectors so that it is efficient in storage [16]. SVM itself is use to find the best hyperplane by maximizing the distance between classes where hyperplanes as illustrated in Figure 2.7 have a meaning or function for separation between classes while in 3D the function used for classification between classes is called plane similarly while in 2D the function used for classification between classes is referred to as line whereas and the mention of separation between classes for higher dimensions is called hyperplane.

The SVM algorithm is often used for text classification or image recognition and basically the steps in this algorithm work, namely first determining class boundaries or hyperplanes and then proceeding with selecting hyperplanes based on maximum margins. The support vector that has been mentioned means that the data point is closest to the hyperplane and can affect the orientation of the hyperplane. By using support vectors, we maximize margins and if the support vector is removed, the position of the hyperplane will also change [17].

2.5 Natural Language Processing (NLP)
Natural language processing (NLP) is a branch of computer science, namely in the extension of artificial intelligence, or what is commonly called AI, which is concerned with giving computers the

ability to understand text and spoken words in the same way that humans can. NLP itself combines computational linguistics, namely rule-based modeling of human language using statistical ML and deep learning models so that simultaneously this model is capable to process human language in the form of text or sound and is able to understand it, including the author's intentions and the author's sentiment The language used by humans is very ambiguous and full of sarcasm or hidden meaning behind it, so it is very difficult to make software or software that is able to accurately read the intended meaning of a word, but programmers must be able to teach natural language-based applications to recognize or understand. the meaning of a word or sound accurately. The following are some of the tasks parsed from the NLP process [18] :

✓ Speech Recognition: commonly called speech-to-text, commonly found in translation machines such as Google Translate, which converts voice into text form
✓ Part of speech tagging: also called grammatical tagging this process usually includes determining the speech part of a word or part of a particular text based on context or usage
✓ Word Sense disambiguation: namely choosing the right meaning for a word that has many meanings and matching it with the context in which the word is used, this is done through a process of semantic analysis
✓ Named entity recognition: identify words or phrases as useful entities.
✓ Co-reference resolution: the task of identifying if and when a text can mean the same
✓ Sentiment Analysis: an attempt to extract sentiment, emotion, confusion from a text
✓ Natural Language generation: the opposite of voice recognition or speech-to-text; task of putting structured information into human language.

2.6 CRISP DM
CRISP-DM stands for Cross Industry Standard Process for Data Mining which is included in one of the models of Data Mining. Built-in 1996 by five companies, namely Integral Solution Ltd (ISL), Daimler AG, Teradata, OHRA, and NCR Corporation and then CRISP-DM was developed and used by many companies in Europe as a non-proprietary standard methodology for data mining. Then between 2006 and 2008 with a desire to update the model CRISP-DM 2.0 SIG was formed. In the CRISP-DM method, six stages are passed as a depiction of the data mining project life cycle, namely Business Understanding, Data

Understanding, Data Preparation, Modeling, Evaluation, and Deployment [19].

3.   METHODOLOGY
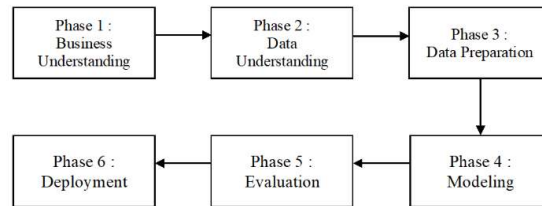
Figure 6. Research Stages

This research will use the CRISP-DM methodology and follow the phases contained in the CRISP-DM (Cross Industry Standard Process for Data Mining) method in which sentiment analysis is carried out following the six phases in the CRISP-DM method, namely from business understanding, data Understanding, data preparation, modeling, evaluation to deployment. Figure 7 is the framework for this research. CRISP-DM is a framework used in data mining projects this research starts by determining the Business understanding of the data mining project in which to help organizations to understand public sentiment towards investment topics in Indonesia and then this research determined Data understanding to understand the type of data that will be crawled and collected. The third phase in data preparation focuses more on data cleansing with several methods that will be described in point 3.3 to make data ready for analysis. The fourth phase is about modeling where this research used the Naïve Bayes classifier and Support Vector Machine algorithm to predict sentiment. In the fifth phase which is deployment, the model is deployed and run and finally, the final phase is Evaluation to evaluate the accuracy and consistency of a prediction.

3.1. Business Understanding
In this phase four tasks need to be done, namely:
1. Determine business objectives: this phase is to understand the goals to be achieved in the analysis of investment topic sentiment in Indonesia so that insight can be found to design a strategy for investment development in Indonesia
2. Assess Situation: determine the availability of sources where in this study the researchers used sources in data processing that were still free.
3. Determine data mining goals: the goals of data mining project is determined in this phase, namely to be used in the sentiment analysis process.
4. Produce project plan / Plan activities: choose the technology to be used and define the plan for each of the existing phases.

In this phase, data from Twitter with certain keywords will be retrieved to get information on investment topics in Indonesia and the information data will be carried out sentiment analysis which is the subject of research.

3.2. Data Understanding

The next phase is the Data Understanding phase where at this stage four tasks need to be carried out, namely:

1. Collect initial data: at this stage we collect data that will be needed in our analysis, namely taking tweets with certain keywords for data on Twitter tweets that discuss investment.

2. Describe the data: examine the data and describe the data used, namely Twitter data in the form of tweets drawn with predetermined keywords.

3. Explore data: the data is processed and several exploration techniques are carried out on the data.

4. Verify data quality: check the quality of the data and document any issues contained in the data.

Data collection was carried out using text-mining techniques using a programming language python and use the tweepy library to take advantage of the free Twitter developer account. The keywords used in data retrieval are as follows:

1. Keyword query search : "Invest OR invest AND (-filter:retweets AND -filter:replies)"

2. Keyword query search : "shares AND (-filter:retweets AND -filter:replies)"

3. Keyword query search : "Bonds OR (bonds AND gas) AND (-filter:retweets AND -filter:replies)"

4. Keyword query search : "Reksadana OR (mutual AND funds) AND (-filter:retweets AND -filter:replies)"

5. Keyword query search : "Deposito OR (depo AND sito) AND (-filter:retweets AND -filter:replies)".

3.3. Data Preparation

In this phase the data is prepared in such a way that it can be used for analysis in which there is a process of selecting data, cleaning data, building data, and integrating data to format data according to analysis needs.

1. Removing unnecessary data: at this stage, the data is cleaned by deleting unnecessary data, both columns and duplicate data.

2. Data Labeling: namely the process of labeling data information to be positive, and negative where positive labels are marked with the number 1 and negative are marked with the number 0.

3. Data Cleansing: this process includes determining whether data is incorrect or incomplete or inappropriate and making some adjustments to the data so that the data quality can be better. Then there

is the text preprocessing stage that is carried out, namely:

- ✓ Case Folding: the benefit of this stage is to generalize the use of capital letters, for example, if the word is in the form of "SHAAM INVESTMENT", then case folding changes all letters to lowercase to "stock investment".

- ✓ Tokenizing: at this stage, the tweet data in the form of sentences will be broken up into pieces called tokens in Indonesian we can refer to the literary module.

- ✓ Filtering: At this stage, important words will be taken as a result of the tokenizing process, and conjunctions that have no meaning which are commonly called stopwords will be removed.

- ✓ Stemming: this stage is the stage needed to reduce the number of different indices of data so that the word will be re-formed

- ✓ Early in this process, for example, the word "using" will change to "use".

3.4. Modelling

In this phase, model development will be carried out with various modeling techniques. In this study, researchers used the Naïve Bayes classifier algorithm, Support Vector Machine (SVM), Decision Tree, Random Forest, and Gradient Boost. with the initial objective. When modeling with a predetermined algorithm the data will be separated into testing data and training data the process will be carried out with K-fold cross-validation and the process will be continued with the TF-IDF (Term Frequency-Inverse Document Frequency) method this process is carried out by giving repeated weights to word in the float value to be used by the classifier. Then the last of this process will be evaluated using a confusion matrix consisting of accuracy, precision, recall, and f-measure.

3.5. Evaluation

At the evaluation phase, there will be a process where the existing models will be compared to another existing model The results will be re-evaluated and re-analyzed then in this stage the next step will also be to determine whether to re-model or proceed to the next stage, namely deployment for the results of the analysis that has been carried out.

3.6. Deployment

This stage is the last stage or phase of the CRISP-DM life cycle, namely in this phase the application of the model that has been produced for business and strategy determination on investment topics in Indonesia will be carried out and becomes a source of information for strategic planning for a business.

4.   RESULT & DISCUSSION

4.1. Evaluation

This study uses 25,000 tweets from Twitter using two algorithms, namely Naive Bayes Classifier and Support Vector Machine. This is done to find out which algorithm has the highest level of accuracy and comparing the results of these two algorithms. The data taken has gone through the pre-processing stage which includes stop-words removal, case folding, removing duplication, stemming and removing unnecessary symbols.

4.1.1.   Naïve Bayes Classifier

Naïve Bayes classifier is a machine learning Naïve Bayes classifier is a machine learning algorithm that learns patterns in data with a supervised method. Naïve Bayes classifier is often used in a classification task like text classification or Natural language processing (NLP). It is also best known for its probabilistic classifier as it is based on the Bayes theorem. In this model, we use the Naïve Bayes classifier as an algorithm to predict the outcome of public sentiment from a tweet that's been crawled using Twitter Developer API. The result shows that the Naïve Bayes Classifier is capable to predict or differentiate positive and negative sentiment with an accuracy rate of 94%.

4.1.2.   Support Vector Machine

Support vector machine is an algorithm that is best known for its advanced classification and regression techniques. The model works by improving prediction accuracy without overfitting trained data. Support Vector Machine has a strong performance in working with large data. This research uses over 20,000 pieces of data which can be classified as large data. Using this algorithm in this research can create a model that is high in accuracy to prediction. The model shows a result of 95% accuracy which can be expected from this model because this model is a good fit for a project that uses a huge amount of data. SVM works better to predict a sentiment than Naïve Bayes classifier by 1% higher.

4.1.3.   Comparison of SVM vs Naïve Bayes Classifier model

In order to evaluate the performance of the machine learning model this study uses a confusion matrix f1 score, accuracy, precision, and recall. Each metric has a different formula and the following table 1 will show a comparison of each metric for both machine learning algorithm.

Model	Accuracy	Precision	Recall	F-1 Score
SVM	95%	96%	96%	96%
Naïve Bayes	94%	91%	91%	89%

Table 1. Comparison of Confusion Matrix Between Model

4.2. Deployment

Based on the results of an analysis of investment topics in Indonesia using 25,000 tweets that have gone through the pre-processing stage where the data was mined from September to November 2022 found words related to the topic of investment in Indonesia where this word often appears with the keyword "invest" so that it can be a reference for organizations to determine steps what they have to do to respond to public sentiment. Below are words related to positive sentiment for the topic "invest" in Indonesia presented in word-cloud form using the python programming language.

Figure 7. Wordcloud Positive Sentiment – Investment

In the word cloud we can see top words that are related to investment topics in Indonesia these words are often time associated with investment starting from "investasi nasional" which translated to "national investment", "kawal investasi" which translated to "guard investment", "wujudkan pemulihan ekonomi" which translated to "achieve economy recovery", investasi modal" which translated to "capital investment" and "selamatkan Indonesia" which translated to "save Indonesia" the bigger you see the word appear in the above picture the more often the word appear in the tweet.

We can also capture the negative sentiment associated with investment topic in Indonesia by using the word-cloud library and from that we can see on topics that are often times associated with "investment" keyword that has a negative emotion presented in the sentence (tweet). Down below picture is a word-cloud for investment topic in Indonesia in negative sentiment.

Figure 8. Word-cloud Negative Sentiment – Investment

In Figure 9 we can see the word cloud for negative sentiment starting from the biggest cloud "Investasi bodong" which translated to "bogus investment", "rugi" which translated to "loss", "tipu" which translated to "con", korban investasi" which translated to "investment victim". These words can give an overall picture of issues occurred in investment topic in Indonesia so that business or organizations can use this as a reference to build their strategy to answers the negative sentiment. A study in 2019 also conducted research to read a sentiment of sentiment analysis by scrapping tweets to capture the sentiment of presidential candidates in Indonesia. The research was also done by several data preparation methods such as text processing, testing, and training data. This study uses a Naïve Bayes algorithm and managed to predict that Jokowi-Ma'ruf amin was positive at 45,45% and negative value of 54,55%. The Prabowo-Sandiaga pair was positive sentiments score of 44,32% and a negative score of 55,68%. The combined data show an accuracy score of 80,1% [20].

5. CONCLUSION

This study aims to use sentiment analysis as a tool to capture people's emotions or sentiments towards the concept of investment in Indonesia with the hope that it can become a measuring tool or reference in developing investment in Indonesia as a large country (population of more than two hundred million people). this research uses Twitter as the main data source by using the keyword "invest" with the data crawling method. after going through several stages, it was found that the developed model produced an accuracy of 95.7% for Support Vector Machine Model and a level of accuracy of 94.6% for the Naive Bayes classifier. it can be concluded that the accuracy results are almost similar but SVM works more superior in sentiment classification. However, there is still some room for improvement in this study. the algorithm used in this study is a machine learning algorithm used to predict an outcome to capture the sentiment of a tweet. a researcher should consider using deep learning algorithm to predict outcome as deep learning have a similar way of processing information as human. we can use BERT as one of many deep learning algorithms since BERT works best in NLP task. Recommendations for future research would be better if sentiment analysis used data from various sources not only Twitter but can combine various social media data so that sentiment analysis can produce a comprehensive and more accurate picture. research can combine data from Twitter, Instagram, Youtube, TikTok, and google reviews. It is also hoped that the model can be developed into an application that can be used by anyone to monitor various topics of discussion and issues. if a model can be applied in a company or organization then the model can be used more flexibly by various users and used to monitor various topics. In this study, there are still many improvements that can be made, namely the amount of data that can be reproduced and queries that can be developed so that the CRISP-DM process can run more effectively. then this research can use other model algorithms to get more accurate model results.

REFERENCES:

[1]   E. Rosenberg, "What is an Investment? - Investments Explained," *https://www.thebalance.com/what-is-an-investment-5090542*, 2021.

[2]   HSBC, "Popular Types of Investment in Indonesia (Jenis-Jenis Investasi yang Populer di Indonesia)," *https://www.hsbc.co.id/1/PA_esf-ca-app-content/content/indonesia/personal/offers/news-and-lifestyle/files/articles/html/201906/jenis-jenis-investasi-yang-populer-di-indonesia.html*, 2019.

[3]   S. (CNBC I. Sidik, "Thanks to the Pandemic, People Become Literate for Investment (Berkah Pandemi, Masyarakat Jadi Melek buat Investasi)," *https://www.cnbcindonesia.com/market/20211123145137-17-293752/berkah-pandemi-masyarakat-jadi-melek-buat-investasi*, Nov. 23, 2021.

[4]   Sony, "Millennial Generation, The Main Factor for Increasing Retail Investors in the Capital Market (Generasi Milenial, Faktor Utama Peningkatan Investor Retail di Pasar Modal)," *https://feb.ugm.ac.id/id/berita/3370-generasi-milenial-faktor-utama-peningkatan-investor-retail-di-pasar-modal*, Oct. 14, 2021.

[5] M. I. Mahdi, "Social Media Users in Indonesia Reach 191 Million in 2022 (Pengguna Media Sosial di Indonesia Capai 191 Juta pada 2022)," *https://dataindonesia.id/digital/detail/peng guna-media-sosial-di-indonesia-capai-191-juta-pada-2022*, Feb. 25, 2022.

[6] V. A. Dihni, "These are the 10 countries with the most Twitter users, is there Indonesia? (Inilah 10 Negara dengan Pengguna Twitter Terbanyak, Ada Indonesia ?)," *https://databoks.katadata.co.id/datapublish /2021/11/04/inilah-10-negara-dengan-pengguna-twitter-terbanyak-ada-indonesia*, Nov. 04, 2021.

[7] F. Amaliah and I. K. Dwi Nuryana, "Comparison of the Accuracy of the Lexicon Based and Naive Bayes Classifier Methods in the Analysis of Public Opinion Sentiment on Investment Applications on Twitter Media (Perbandingan Akurasi Metode Lexicon Based Dan Naive Bayes Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter)," *Journal of Informatics and Computer Science (JINACS)*, vol. 3, no. 03, pp. 384–393, Apr. 2022, doi: 10.26740/jinacs.v3n03.p384-393.

[8] N. Helmiah *et al.*, "Application of the Naïve Bayes Method in the Analysis of Public Perceptions regarding the Plan to Ratify the 2020 Omnibus Law Bill in the Investment and Employment Sector in Indonesia (Penerapan Metode Naïve Bayes dalam Analisis Persepsi Masyarakat mengenai Rencana Pengesahan RUU Omnibus Law di Bidang Investasi dan Ketenagakerjaan Tahun 2020 di Indonesia)," *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, vol. 8, no. 2, p. 48, Dec. 2020, doi: 10.24252/msa.v8i2.16743.

[9] A. D. Adhi Putra, "Sentiment Analysis on User Reviews of the Seeds and Bareksa Applications with the KNN Algorithm (Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma) KNN," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 8, no. 2, pp. 636–646, Jun. 2021, doi: 10.35957/jatisi.v8i2.962.

[10] B. M. Hasibuan, "Investment and The History of Foreign Investment in Indonesia (Investasi dan Sejarah Perkembangan Investasi asing di Indonesia)," *https://business-law.binus.ac.id/2017/02/19/investasi-dan-sejarah-perkembangan-investasi-asing-di-indonesia/#_ednref1*, Feb. 2017.

[11] Twitter, "New user FAQ," *https://help.twitter.com/en/resources/new-user-faq*, 2022.

[12] C. Mutia Annur, "The number of daily active Twitter users continues to increase until the first quarter of 2022 (Jumlah Pengguna Aktif Harian Twitter Terus Naik sampai Kuartal I 2022)," *https://databoks.katadata.co.id/datapublish /2022/07/13/jumlah-pengguna-aktif-harian-twitter-terus-naik-sampai-kuartal-i-2022*, 2022.

[13] S. Gupta, "Sentiment Analysis: Concept, Analysis and Applications," *https://databoks.katadata.co.id/datapublish /2022/07/13/jumlah-pengguna-aktif-harian-twitter-terus-naik-sampai-kuartal-i-2022*, 2018.

[14] D. Johnson, "Unsupervised Machine Learning: Algorithms, Types with Example," *https://www.guru99.com/unsupervised-machine-learning.html*, 2022.

[15] R. Gandhi, "Naive Bayes Classifier," *https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c*, 2018.

[16] Scikit-learn, "1.4. Support Vector Machines," *https://scikit-learn.org/stable/modules/svm.html*, 2022.

[17] S. Hussein, "Support Vector Machine, Algoritma untuk Machine Learning," *https://geospasialis.com/support-vector-machine/*, 2021.

[18] IBM, "Natural Language Processing (NLP)," *https://www.ibm.com/cloud/learn/natural-language-processing*, 2020.

[19] T. Mauritsius, "Cross-Industry Standard Process for Data Mining (CRISP-DM)," *https://mmsi.binus.ac.id/2020/09/18/cross-industry-standard-process-for-data-mining-crisp-dm/*, 2020.

[20] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Oct. 2019, pp. 1–5. doi: 10.1109/ICIC47613.2019.8985884.