JURNAL SISFOTENIKA
Vol. 13 No. 2, Juli 2023
p-ISSN: 2087-7897 ; e-ISSN : 2460-5344
DOI : 10.30700/jst.v13i2.1397

Algoritma Pohon Keputusan pada Analisis Sentimen Terhadap Perang Rusia dan Ukraina

The Decision Tree Algorithm on Sentiment Analysis: Russia and Ukraine War

Apriandy Angdresey*[1], Geraldo Saroinsong[2]
[1,2] Department of Informatics Engineering, Universitas Katolik De La Salle, Manado
e-mail: *[1]**aangdresey@unikadelasalle.ac.id**, [2]geraldosaroinsong5@gmail.com

Abstrak

Analisis sentimen merupakan suatu metode yang secara otomatis memahami, mengekstrak, dan mengolah suatu data teks untuk memperoleh suatu informasi atau pendapat yang terkandung dalam data tersebut. Hal ini dilakukan untuk menentukan suatu pendapat, baik itu positif, negatif ataupun netral dalam relevansinya dengan suatu objek, produk atau topik. Pada penelitian ini bertujuan untuk mengimplementasikan algoritma pohon keputusan ke dalam aplikasi analisis sentimen yang terkait pada perang Rusia dan Ukraina. Isu ini telah menjadi trending topik di twitter dan menarik banyak perhatian dari masyarakat nasional maupun internasional. Oleh karena itu, penulis mengumpulkan pendapat atau komentar dan cuitan dari pengguna twitter, dan menganalisisnya menggunakan algoritma pohon keputusan. Penelitian ini, penulis berhasil mengumpulkan 1.069 data dan mendapatkan hasil dengan nilai rata-ratanya. Pada penelitian ini melakukan 15 pengujian dengan menggunakan tiga partisi data yang berbeda, yaitu 80:20, 70:30, dan 60:40. Berdasarkan hasil yang diperoleh penulis menyimpulkan bahwa partisi data 80% data pelatihan dan 20% data pengujian memperoleh skor rata-rata tertinggi, dengan hasilnya yaitu 85,61% untuk akurasi, 86,27% untuk presisi, dan untuk nilai *recall* sebesar 86,01%.

Kata Kunci: Analisis Sentimen, Klasifikasi, Pohon Keputusan, Twitter

Abstract

Sentiment analysis is a method to automatically understand, extract, and process a text data to obtain an information or an opinion contained in said data. This is done in order to determine whether an opinion is positive, negative or neutral in relevance to an object, product or topic. This research is aimed to implement the decision tree algorithm into our sentiment analysis application regarding the Russian and Ukraine war. This issue has been a trending topic on twitter and is attracting many attention from the public. Therefore, we gathered their opinions and analyze them using the decision tree algorithm. We managed to gather 1.069 data and obtain the results along with their average scores by conducting 15 tests using three different data partitions. Based on the results obtained we concluded that the data partition of 80% training data and 20% testing data gained the highest average score. The result was 85.61% for accuracy, 86.27% for the precision, and 86.01% for the recall.

Keywords**—***Sentiment analysis, classification, Decision Tree, Twitter*

1. INTRODUCTION

Sentiment analysis is a method used to process various opinions given by users, both experts and consumers through various media about a product, service, agency, or topic. Sentiment analysis is a method to understand, extract and process textual data automatically to find positive, negative, and neutral opinions. The data used comes from various social media. The purpose of this method is to understand the subjectivity of text data and determine the bias of an opinion towards a problem or an object which is achieved by transforming text data into numbers that can be calculated. These numbers can then be used to find out whether an opinion is positive or negative.

The Decision Tree algorithm is a classification method where an attribute is represented as a node (the first node is called the root node) and the value as branches that form a tree. The decision tree algorithm processes data by changing its form into a tree model which can result in a set of rules. The decision tree algorithm requires less effort to prepare data in the pre-processing stage compared to other algorithms. However, small changes in the data can result in a big transformation in the tree model which can cause it to be unstable. Our goal in this study is to implement the decision tree algorithm into a sentiment analysis application to analyze the public sentiment regarding the Russia and Ukraine war by using the data are gathered from Twitter.

Twitter is one of the most used social media applications. Its users utilize it to bring forward their opinions on a tweet about various topics using hashtags. Twitter is considered a social media due to the users being able to have friends, like, share, and re-post (retweet) other users posts. As time develops Twitter is becoming a place to report news on. One of the trending topics today is the war between Russia and Ukraine which began on 24 February 2022. The war attracted much attention from everyone around the globe. The purpose of this research is to analyze the many sentiments on the Russia and Ukraine war using a decision tree algorithm. Specifically, our contribution is the application can assist journalists in collecting information about public opinion towards the war.

The remainder of this paper is organized as follows: Section II explains the related works of this study, and Section III has presented the method used in this study which are text mining, decision tree, and the confusion matrix. Further, in Section IV elaborated in detail on the results and discussed this study, and the conclusion of this study is shown in Section VI.

2. RELATED WORKS

Decision Tree is a decision analysis technique included in the classification method, which can help decision-makers when faced with several choices by projecting possible outcomes. In addition, it can also show the possible factors that will affect the decision alternatives, accompanied by an estimate of the final results that will be obtained when taking the decision alternative [1]. The use of decision trees is the ability to simplify complex decision-making processes into simpler ones so that decision-makers will more easily interpret solutions to problems. There are several advantages of using a decision tree, namely eliminating unnecessary calculations, and the sample being tested is only based on certain criteria or classes. Moreover, the area of decision-making that was previously complex and highly global, can be changed to become more simple and specific [2].

In multivariate analysis, with a large number of criteria and classes, it is usually necessary to estimate either the high dimensional distribution or certain parameters of the class distribution. Furthermore, the decision tree method avoids the emergence of problems by using fewer criteria at each internal node without significantly reducing the quality of the resulting decisions. In addition, it is flexible, choosing features from different internal nodes, the selected features will distinguish a criterion from other criteria in the same node [3]. The flexibility of this decision tree method improves the quality of the resulting decisions when compared to using the more conventional one-stage calculation method. As in the study [4], the authors predict the swimming pool water quality using the decision tree of the Iterative Dichotomiser 3 algorithm, with an accuracy rate is 100% and a statistical value of kappa 1. In addition, in analyzing the effect of stock prices on the financial fundamentals of a company, the authors build stock investments using a decision tree model through the ID3 algorithm, the results of this study are to provide decision support for future investments through analysis of the current financial and market situation [5].

Text Mining is a process of extracting information or pattern extraction in the form of useful information and knowledge from a large number of text data sources, such as Word documents, PDFs, text quotes, etc [6]. The result of this process is obtained for the certain purposes. Sentiment analysis is one part of text mining that has begun to be widely used in various fields. Sentiment analysis is often also referred to as opinion mining which is carried out to analyze large amounts of data using several algorithms. This shows that it will dig into the emotions of every customer’s comments or sentences. Nowadays, the customers are very happy to express their feelings through online platforms, such as social media, e-commerce, and websites. Therefore, sentiment analysis is carried out on these platforms, for example in politics [7], products marketing [8], services [9], hospitality [10], and others. One of the advantages of sentiment analysis is that it saves time and effort.

3. METHOD

This section will elaborate the steps we used to process the data. Suppose we have gathered text data as shown in Table 1. These data will go through three stages, which consists of pre-processing, processing and post-processing. Table 2 is the result of the pre-processing stage, which are cleaning, case folding, stopword filtering, stemming and tokenizing.

Table I. Raw Data

No.	Tweet	Label	Type
1.	@jamesEwoo:This war is sad :(#UkrainwRussiaWar	Positive	Training
2.	@willTennyson: Hail Russia BURN Ukraine #UkraineRussiaWar	Negative	Training
3.	@jake420: Invasion is not the play #UkraineRussiaWar	Positive	Training
4.	@TheTerminator: I am supporting Russia in this war!!! #UkraineRussiaWar	Negative	Training
5.	@BigChungus: We are longinng for Peace in Russia and Ukraine #UkraineRussiaWar	Neutral	Training
6.	This invasion is NOT looking good for both countries!!! No more war!!!	?	Testing

After going through the pre-processing stage the data will go through the processing stage where the number values will be processed using the decision tree. We need to calculate the Entropy and Gain values of each terms in Table 3 using the formulas shown below. To have a better understanding on how we calculated the values, an example has been included with the term “*war*” which can be seen below. The result of Entropy and Gain can be seen on Table 4.

$$Entropy(S) = \sum_{i=1}^{n} P_i \times log_2(P_i) \tag{1}$$

Table 2. Data Pre-Processing

No.	Tweet	Pre-Processing
1.	@jamesEwoo:This war is sad :(#UkrainwRussiaWar	war, sad
2.	@willTennyson: Hail Russia BURN Ukraine #UkraineRussiaWar	hail, russia, burn, ukraine
3.	@jake420: Invasion is not the play #UkraineRussiaWar	invasion, not, play
4.	@TheTerminator: I am supporting Russia in this war!!! #UkraineRussiaWar	i, support, russia, war
5.	@BigChungus: We are longinng for Peace in Russia and Ukraine #UkraineRussiaWar	we, long, peace, russia, ukraine
6.	This invasion is NOT looking good for both countries!!! No more war!!!	war, not, good, both, country, more, war

$$Gain(A) = Entropy(total) - Entropy(class) \tag{2}$$

Table 3. TF-IDF

Term	1	2	3	4	5	DF
war	1	0	0	1	0	2
sad	1	0	0	0	0	1
hail	0	1	0	0	0	1
russia	0	1	0	1	1	3
invasion	0	0	1	0	0	1
not	0	0	1	0	0	1
play	0	0	1	0	0	1
i	0	0	0	1	0	1
support	0	0	0	1	0	1
we	0	0	0	0	1	1
long	0	0	0	0	1	1
peace	0	0	0	0	1	1
ukraine	0	0	0	0	1	1

For example, word ”*war*”, we can calculate the entropy for the positive class as *Entropy*(+) = −1/5 ∗ log_2 1/5 = 0.4642, whereas for the *Entropy*(−) = −1/5 ∗ log_2 1/5 = 0.4642, and *Entropy*(*n*) = 0. Furthermore, the gain value, *i.e. Gain* = 1.52 + 0.9284 = 2.4484.

Table 4. Entropy and Gain Values

Terms	Entropy(+)	Entropy(*n*)	Entropy(-)	Gain
war	0.4642	-	0.4642	2.4484
sad	0.4642	-	-	1.9842
hail	-	-	0.4642	1.9842
russia	-	0.5284	0.4642	2.5126

invasion	0.4642	-	-	1.9842
not	0.4642	-	-	1.9842
play	0.4642	-	-	1.9842
i	-	-	0.4642	1.9842
support	-	-	0.4642	1.9842
we	-	0.4642	-	1.9842
long	-	0.4642	-	1.9842
peace	-	0.4642	-	1.9842
ukraine	-	0.4642	0.4642	2.4484

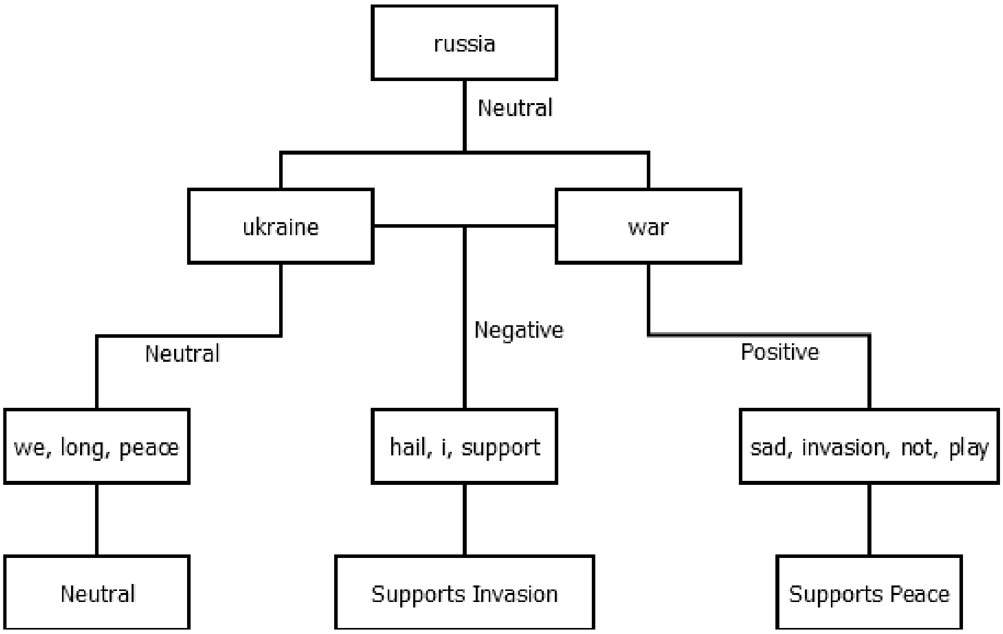

Figure 1. The Decision Tree Model

Based on the results obtained, a tree model can be generated which is shown on Fig. 1. Furthermore, a set of rules can be generated based on said tree model which is shown below:
(*russia*) & (*ukraine*) & (*we, long, peace*) = Neutral
(*russia*) & (*war*) & (*sad, invasion, not, play*) = Positive(Supports Peace)
(*russia*) & (*Ukraine,war*) & (*hail, i, support*) = Negative(Supports Invasion)

3. RESULTS AND DISCUSSION

This section will discuss the result of the research and it was achieved. The flowchart of the application has been included in this paper which can be seen in Figure 2 to give a better understanding of how the application operates. The process of the application is divided into three main parts. The first part is pre-processing. In this part, the user will enter a data count for the application to fetch. After fetching said data the user will label each data accordingly in order to train the algorithm.

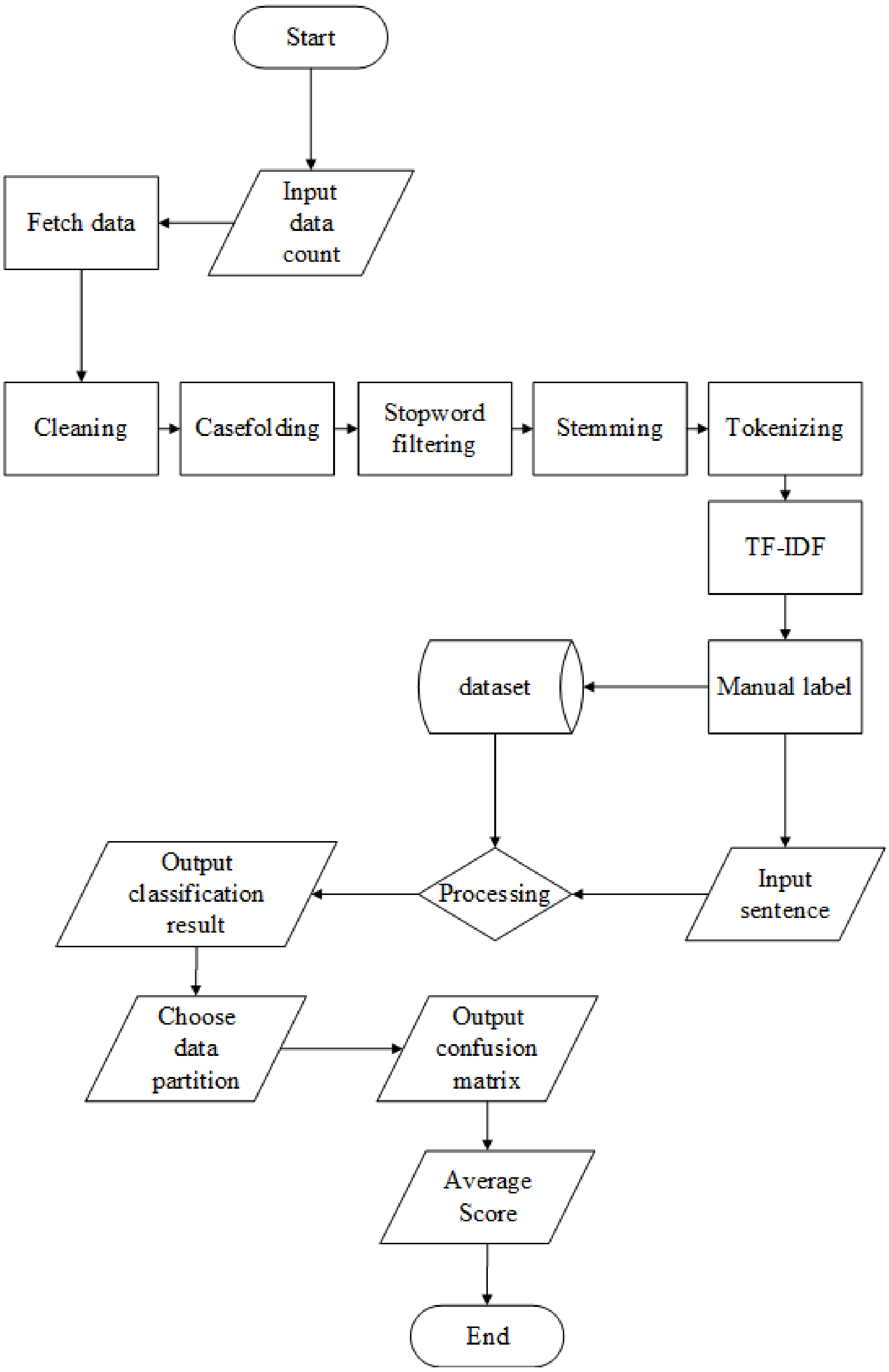

Figure 2. Application Flowchart

Text data needs to be cleaned and encoded to a numeric value before being fed to the decision tree method, this process of cleaning and encoding is known as text preprocessing. This application uses text preprocessing, namely cleaning, case-folding, stop-word filtering, stemming, tokenizing, and TF-IDF.

In these stages, the data will be cleaned by removing symbols, Twitter accounts, hashtags, and words that have no significance. Furthermore, these text data will be converted into numbers. At the case folding stage, it is one of the simplest and most effective forms of text preprocessing, although it is often ignored, which aims to convert all letters in the document into lowercase, characters other than letters are removed changing text to lowercase, removing punctuation marks, and removing whitespace. Characters other than letters are omitted and considered delimiters. Tokenizing is the process of separating the text into pieces called tokens for later analysis. Words, numbers, symbols, punctuation marks, and other important entities can be considered tokens. In NLP, tokens are interpreted as "words" although tokenization can also be done in paragraphs or sentences. Filtering is the stage of taking important words from the token results by removing less important words or wordlists to save important words. Stopwords are common words that usually appear in large numbers and are considered meaningless. The meaning behind using stopwords is that by removing words that have low information from a text, we can focus on important words instead. Meanwhile, Stemming is the process of removing the inflection of a word into its base form, however, the basic form does not mean the same as the root word.

The application will then put the data through a few stages, namely cleaning, case-folding, stop-word filtering, stemming, tokenizing and TF-IDF. In these stages the data will be cleaned by removing symbols, twitter handles, hashtags and words that have no significant meaning. Furthermore these text data will be converted into numbers. The second main part is processing in which the data will be calculated using the decision tree algorithm to determine their respective sentiment levels. This is achieved by calculating the Entropy of each class and their Gain values. The user then may move in to the next part which is post processing where they can choose a data partition in order to test the accuracy of the algorithm. The application will then use all of the data gathered so far and divide them into two parts according to the data partition and display the result in the form of a confusion matrix. The user also can view the average score of accuracy, precision and recall of every tests done thus far.

We managed to gather 1,069 tweet data and ran some tests using the three data partitions in the application, which are 80-20, 70-30 and 60-40. We ran 15 tests for each data partition and calculated their average score. After running the tests we managed to obtain the average score of each data partition with the 80-20 data partition having the highest score. The test results obtained along with the average scores have been included in this paper which can be viewed in Table 5, 6, and Table 7.

Table 5. 80-20 Data Partition

Test	Execution Time	Accuracy	Precision	Recall	F1-Score
1	0.22 sec	87.38%	87.56%	87.67%	87%
2	0.21 sec	83.18%	83.59%	83.21%	83%
3	0.23 sec	85.05%	84.84%	85.2%	85%
4	0.23 sec	83.18%	83.74%	83.47%	83%
5	0.19 sec	86.92%	86.86%	87.38%	87%
6	0.22 sec	85.05%	87.07%	84.77%	85%
7	0.23 sec	85.98%	86.39%	85.94%	86%
8	0.25 sec	80.84%	82.25%	80.8%	81%
9	0.22 sec	85.05%	86.11%	85.02%	85%
10	0.19 sec	82.24%	83.57%	82.58%	82%
11	0.2 sec	92.06%	93.06%	92.11%	92%
12	0.21 sec	87.85%	88.1%	88.2%	88%
13	0.2 sec	84.11%	84.63%	84.38%	84%
14	0.22 sec	87.85%	88.36%	87.79%	88%
15	1.25 sec	87.38%	87.94%	88.08%	87%
Average	**0.28 sec**	**85.61%**	**86.27%**	**86.01%**	**85.71%**

Table 6. 70-30 Data Partition

Test	Execution Time	Accuracy	Precision	Recall	F1-Score
1	0.57 sec	81.62%	82.28%	81.4%	82%
2	0.23 sec	79.13%	79.56%	79.49%	79%
3	0.22 sec	75.08%	76.9%	75.67%	75%
4	0.22 sec	83.8%	83.88%	83.74%	84%
5	0.2 sec	79.13%	81.08%	79.36%	79%
6	0.22 sec	75.08%	75.88%	75.14%	75%
7	0.21 sec	75.7%	77.26%	76.32%	76%
8	0.25 sec	78.5%	79.72%	78.81%	78%
9	0.23 sec	78.5%	79.41%	78.4%	78%
10	0.24 sec	79.95%	77.23%	77.03%	77%
11	0.24 sec	79.44%	79.88%	79.51%	79%
12	0.21 sec	80.69%	81.66%	80.45%	81%
13	0.27 sec	80.37%	81.49%	80.05%	80%
14	0.23 sec	78.5%	78.84%	78.41%	78%
15	0.18 sec	77.88%	77.89%	77.83%	78%
Average	**0.24 sec**	**78.89%**	**79.53%**	**78.77%**	**78.60%**

Table 7. 60-40 Data Partition

Test	Execution Time	Accuracy	Precision	Recall	F1-Score
1	0.19 sec	71.73%	73.01%	72.87%	72%
2	0.21 sec	71.96%	74.07%	72.42%	72%
3	0.23 sec	73.6%	74.96%	73.29%	73%
4	0.23 sec	72.9%	77.08%	73.67%	73%
5	0.19 sec	71.5%	71.81%	71.65%	71%
6	0.23 sec	73.6%	74.76%	74.03%	74%
7	0.22 sec	72.43%	73.6%	72.47%	72%
8	0.19 sec	70.56%	71.9%	70.67%	71%
9	0.23 sec	71.5%	73.95%	71.79%	72%
10	0.22 sec	73.83%	75.28%	74.09%	74%
11	0.2 sec	71.73%	72.75%	71.9%	72%
12	0.18 sec	72.43%	73.99%	72.57%	73%
13	0.2 sec	71.96%	73.87%	72.22%	71%
14	0.2 sec	75.7%	77.85%	75.41%	76%
15	0.2 sec	73.13%	73.39%	73.22%	73%
Average	**0.208 sec**	**72.57%**	**74.15%**	**72.82%**	**72.60%**

4. CONCLUSION

This section will discuss the conclusions of this study, namely the application of a decision tree algorithm in analyzing sentiment towards the Russian and Ukrainian wars has been successfully developed. This application can retrieve a number of tweet data, while for the tests carried out we used a confusion matrix with the best accuracy of 85.61%, 86.27% for the precision, and 86.01% for the recall on a data partition of 80% training data and 20% test data, with an execution time of less than 2 seconds. For future works, we can implement other classification algorithms and make comparisons.

REFERENCES

[1] A.A Supianto, A.J. Dwitama, and M. Hafis. Decision tree usage for student graduation classification: A comparative case study in faculty of computer science brawijaya university. In 2018 International Conference on Sustainable Information Engineering and Technology (SIET), pages 308–311, 2018.

[2] F.J. Yang. An extended idea about decision trees. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI), pages 349–354, 2019.

[3] X.Z. Wang, H.W. Yang, M.H. Zhao, and J. Sun. A decision tree based on hierarchical decomposition. In Proceedings. International Conference on Machine Learning and Cybernetics, volume 4, pages 1824–1828 vol.4, 2002.

[4] A. Angdresey, L. Sitanayah, and V.J.A. Sampul. Monitoring and Predicting Water Quality in Swimming Pools. EPI International Journal of Engineering, 3:119–125, Aug. 2020.

[5] C. Chen. The Apply of ID3 in Stock Analysis. In 6th International Conference on Computer Science Education (ICCSE), pages 24–27,Aug. 2011.

[6] A. Angdresey, M.A. Lamongi, and R. Munir. Information Retrieval System in the Bible. In Cogito Smart Journal, volume 7, pages 111– 120, Jun. 2021.

[7] 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), pages 1–5, 2019.

[8] S. Gusriani, K.D.K.Wardhani, and M.I. Zul. Analisis Sentimen Terhadap Toko Online di Sosial Media Menggunakan Metode Klasifikasi Na¨ıve Bayes (Studi Kasus: Facebook Page BerryBenka). In 4th Applied Business and Engineering Conference, 2016.

[9] A. Angdresey, I.Y. Kairupan, and K.G. Emor. Classification and Sentiment Analysis on Tweets of the Ministry of Health Republic of Indonesia. In 2022 Seventh International Conference on Informatics and Computing (ICIC), pages 1–6, 2022.

[10] E. Indrayun. Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization. Evolusi: Jurnal Sains dan Manajemen, 4:20–27, 2016.